W276–W279 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* *Published online 12 April 2022*
https://doi.org/10.1093/nar/gkac240

Search and sequence analysis tools services from EMBL-EBI in 2022

Fábio Madeira †, Matt Pearce †, Adrian RN Tivey, Prasad Basutkar, Joon Lee, Ossama Edbali, Nandana Madhusoodanan, Anton Kolesnikov and Rodrigo Lopez *

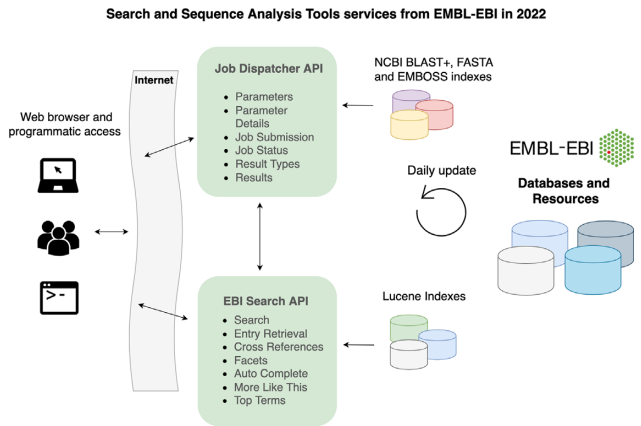
European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received February 04, 2022; Editorial Decision March 28, 2022; Accepted March 28, 2022

ABSTRACT

The EMBL-EBI search and sequence analysis tools frameworks provide integrated access to EMBL-EBI's data resources and core bioinformatics analytical tools. EBI Search (https://www.ebi.ac.uk/ebisearch) provides a full-text search engine across nearly 5 billion entries, while the Job Dispatcher tools framework (https://www.ebi.ac.uk/services) enables the scientific community to perform a diverse range of sequence analysis using popular bioinformatics applications. Both allow users to interact through user-friendly web applications, as well as via RESTful and SOAP-based APIs. Here, we describe recent improvements to these services and updates made to accommodate the increasing data requirements during the COVID-19 pandemic.

GRAPHICAL ABSTRACT



INTRODUCTION

The COVID-19 pandemic caused by the severe acute respiratory syndrome cCoronavirus 2 (SARS-CoV-2) and the lockdown measures implemented by governments worldwide to contain it has caused unprecedented economic and societal disruption (1). This has highlighted the need for the scientific community to work together to effectively tackle the global COVID-19 health crisis. The European Bioinformatics Institute (EMBL-EBI; https://www.ebi.ac.uk/) alongside many other institutions, have promptly made international cooperation networks to provide researchers and the general public access to trustworthy information. In fact, the pandemic has posed an enormous challenge to biological data resources since 2020, with increasing quantities of data being generated and needing to be made available to users from across the scientific communities (2). EMBL-EBI has contributed to the fight against COVID-19 on several fronts, in particular by helping the development of the European COVID-19 Data Portal (3), which leverages and brings together biomolecular data from a variety of EMBL-EBI's data resources and services to researchers, clinicians and public health professionals. Additionally, the EMBL-EBI has over the last thirteen plus years developed Web Service API-centred frameworks, EBI Search and Job Dispatcher (4), for providing access to (i) a free text search and powerful cross-referencing engine and to (ii) bioinformatics sequence analysis tools, respectively, that provide access to these rich data. The services have worked closely with teams across the EMBL-EBI and further afield to expand the capabilities of the frameworks. In this paper, we overview recent improvements and updates made to the services to accommodate the increasing data requirements during the COVID-19 pandemic.

EBI SEARCH AND JOB DISPATCHER

EBI Search (previously EB-eye) is an Apache Lucene-based search engine platform, providing simple and uniform access to the public biological data resources hosted by EMBL-EBI. These data are spread across 160+ data sets (domains), updated daily. Searches may be carried out using a RESTful API or the EBI Search website, returning result sets including hierarchical facets and cross-references to

*To whom correspondence should be addressed. Tel: +44 1223 494 423; Fax: +44 1223 494 468; Email: rls@ebi.ac.uk
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

© The Author(s) 2022. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

other datasets, allowing links to be followed throughout the available resources. The EBI Search engine provides search functionality to services across EMBL-EBI, including ENA (5), Ensembl Genomes (6), the OMICS DI portal (7) and the COVID-19 Data Portal.

The Job Dispatcher tools framework (JD) provides integrated access to core bioinformatics applications and required biological data. The JD catalogue of tools includes some of the most popular powerhouses in bioinformatics, from sequence similarity search applications, such as NCBI BLAST+ (8) and FASTA (9), multiple sequence alignment and pairwise sequence alignment tools, such as Clustal Omega (10) and Kalign (11), tools for functional annotation and prediction such as InterProScan 5 (12), RNA analysis tools such as R2DT (13), to other sequence analysis utilities. The use of sequence similarity search tools comprises 45 000+ distinct sequence libraries from major database resources hosted at EMBL-EBI, including UniProtKB (14), ENA and Ensembl Genomes. The JD framework provides an interface between high-performance compute clusters and command-line applications. Free access to the tools is provided via the service's website, as well as programmatically, via transparent and reliable RESTful and SOAP Web Services APIs. Visual representations of tool results are also provided to help the users understand the job outputs. An additional component of the JD offering is Dbfetch, which provides a common interface to database entry retrieval in a variety of different formats for all the sequence libraries available to search in JD.

UPDATES ON DATA RESOURCES

EBI Search data resources are grouped into a hierarchical tree of domains (see a list of data resources available in EBI Search in Supplementary Table 1). Since the last update, Open Targets (15), VarSite (16), PDBe-KB (17), GWAS Catalog (18), EMPIAR (19), European Variation Archive Studies (20) and Cellosaurus (21) have been added as new resources. Additionally, 18 new COVID-19-specific domains have been added.

New sequence libraries were added to JD (see a list of all the sequence libraries currently provided by JD in Supplementary Table 2), namely SARS-CoV-2 dataset releases from Ensembl, UniProtKB, ENA and Pfam (22). IPD-NHKIR (23) coding and genomic sequences, as well as sequences from the AlphaFold DB (24), EMDB (25) and PDBe-KB, have also been made available for sequence similarity search in the JD tools framework and retrieval via Dbfetch.

UPDATES ON THE TOOLS

Sequence analysis tools running under JD are categorised according to their functionality and have been regularly updated to their latest available versions (see a list of all the categories and bioinformatics tools currently provided by JD in Supplementary Table 3). These were also updated to run in containers with Singularity that future-proof their execution and isolation in an ever-changing computational environment. Three new tools have been added to the framework since the last update, including R2DT, for predicting and visualising RNA secondary structures, SSEARCH2SEQ and GGSEARCH2SEQ, for generating local and global pairwise alignments, respectively (9). A new JSON schema (https://github.com/ebi-wp/sss_json_schema) has been developed enabling a standardised JSON output to be provided for the results of sequence searches by tools in the FASTA and NCBI BLAST+ suites.

USAGE OF THE SERVICES DURING THE COVID-19 PANDEMIC

The EBI Search API has been providing the search results for the COVID-19 Data Portal since that site's inception in April 2020, with the team closely involved with the site's development. Data is retrieved from a combination of domains dedicated to purely COVID-19-related data and wider domains queried with filters applied to limit content. Changes have been made to the indexing process to allow data to be copied between domains, de-normalizing data and improving response times. Deep paging (https://lucidworks.com/post/coming-soon-to-solr-efficient-cursor-based-iteration-of-large-result-sets/) has been enabled in selected COVID-19 domains, enabling search results to be retrieved beyond the one million result limit to assist with data verification. A base web application has been developed from the COVID-19 Data Portal for future EMBL-EBI portal projects specialising in particular areas of health bioinformatics that cut across multiple datasets.

The deployment model and reliability of JD data pipelines for daily indexing of biological databases has been improved. These are now generated using BLAST database version 5 which together with NCBI Taxonomy data (26), enables limiting the search by taxonomy with both inclusion and exclusion lists of NCBI Taxonomy identifiers (TaxIDs). The NCBI Taxonomy database is updated daily, and a tree structure is built using Taxonomy Resolver (https://github.com/ebi-wp/taxonomy-resolver), based on the NCBI Taxonomy Database classification. This functionality is currently available for sequence searches against UniProtKB and ENA databases.

EBI Search and JD are core services used extensively by other resources and portals at the EMBL-EBI and elsewhere. The sheer volume of data being generated during the COVID-19 pandemic has resulted in an average of 2.5 million requests per day to the EBI Search engine, and nearly 500 million analyses being performed under the JD tools framework in 2021. These are well in line with what was observed during 2020, where a noticeable surge of >130 million sequence analysis performed during the COVID-19 outbreak months of April and May 2020, and highlights the importance of the services described here. Trends in tool usage and citation data of our most recent article describing the framework (4) since 2019 highlight how core bioinformatics applications for sequence searching and alignment are fundamental for life sciences research and development, ranging from structural biology and drug discovery to immunology and epidemiology. Among many other use cases, JD tools have been used for aiding: the identification of the SARS-CoV-2 proteome; aiding the analysis of the SARS-

CoV-2 structural and non-structural proteins, including the spike proteins and analysis of their mediated host cell entry; evaluation of new possible drug targets and development of new drugs, vaccine constructs and antibodies; and mutational spectra analysis and detection of new variants (27–31).

DISCUSSION

At the time of writing, the ongoing pandemic is driving us to improve our services to better serve the scientific community and react quickly to global changes in data demand. The EBI Search team will continue to work on the baseline portal web application to make rapid deployment of future portal sites simple. Improving search performance is a priority, and the team will be applying upgrades to the major software libraries used by the platform. The JD team is currently developing a brand new modern and interactive website and backend for the JD tools framework. Work on an updated tabular results page with new interactive features for sorting, selecting and faceting the results, as well as downloading sequences in bulk and initiating workflows, is underway. We hope these changes will further expand the offering of both tools and datasets while maintaining the security, scalability and reliability of the service.

DATA AVAILABILITY

EBI Search is available from https://www.ebi.ac.uk/ebisearch and JD tools are available from https://www.ebi.ac.uk/services. Detailed documentation about how to interface with the services programmatically are provided at https://www.ebi.ac.uk/Tools/webservices. Additionally, users can explore the EBI Search and JD APIs interactively at: https://www.ebi.ac.uk/ebisearch/apidoc.ebi and https://www.ebi.ac.uk/Tools/common/tools/help, respectively. Sample Web Service clients in Python, Perl and Java are also provided for EBI Search as well as JD on the following GitHub repositories: https://github.com/ebi-wp/EBISearch-webservice-clients and https://github.com/ebi-wp/webservice-clients, respectively. CWL command-line tool definitions and example workflows are available from https://github.com/ebi-wp/webservice-cwl. These services are developed in accordance with FAIR principles.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Simone Badoer, Ravi Mahankali, Ijaz Ahmad, Khalid Kamal and Vidya Sankaran Potti and for web administration support. We would like to also thank all EMBL-EBI teams for their invaluable help in providing biological data, applications and expertise.

FUNDING

EMBL-EBI is indebted to its funders, including the EMBL member states and the European Commission through the H2020 Programme under EOSC-Life [824087]; BY-COVID [101046203]; EarlyCause [848158]. Funding for Open Access: EMBL.
Conflict of interest statement. None declared.

REFERENCES

1. Hu,B., Guo,H., Zhou,P. and Shi,Z.-L. (2021) Characteristics of SARS-CoV-2 and COVID-19. *Nat. Rev. Microbiol.*, **19**, 141–154.
2. Cantelli,G., Cochrane,G., Brooksbank,C., McDonagh,E., Flicek,P., McEntyre,J., Birney,E. and Apweiler,R. (2021) The European Bioinformatics Institute: empowering cooperation in response to a global health crisis. *Nucleic Acids Res.*, **49**, D29–D37.
3. Harrison,P.W., Lopez,R., Rahman,N., Allen,S.G., Aslam,R., Buso,N., Cummins,C., Fathy,Y., Felix,E., Glont,M. *et al.* (2021) The COVID-19 Data Portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. *Nucleic Acids Res.*, **49**, W619–W623.
4. Madeira,F., Park,Y.M., Lee,J., Buso,N., Gur,T., Madhusoodanan,N., Basutkar,P., Tivey,A.R.N., Potter,S.C., Finn,R.D. *et al.* (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.*, **47**, W636–W641.
5. Cummins,C., Ahamed,A., Aslam,R., Burgin,J., Devraj,R., Edbali,O., Gupta,D., Harrison,P.W., Haseeb,M., Holt,S. *et al.* (2022) The European Nucleotide Archive in 2021. *Nucleic Acids Res.*, **50**, D106–D110.
6. Howe,K.L., Achuthan,P., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J. *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.
7. Perez-Riverol,Y., Zorin,A., Dass,G., Vu,M.-T., Xu,P., Glont,M., Vizcaíno,J.A., Jarnuczak,A.F., Petryszak,R., Ping,P. *et al.* (2019) Quantifying the impact of public omics data. *Nat. Commun.*, **10**, 3512.
8. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
9. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 2444–2448.
10. Sievers,F. and Higgins,D.G. (2021) The clustal omega multiple alignment package. *Methods Mol. Biol.*, **2231**, 3–16.
11. Lassmann,T. (2019) Kalign 3: multiple sequence alignment of large data sets. *Bioinformatics*, **36**, 1928–1929.
12. Blum,M., Chang,H.-Y., Chuguransky,S., Grego,T., Kandasaamy,S., Mitchell,A., Nuka,G., Paysan-Lafosse,T., Qureshi,M., Raj,S. *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49**, D344–D354.
13. Sweeney,B.A., Hoksza,D., Nawrocki,E.P., Ribas,C.E., Madeira,F., Cannone,J.J., Gutell,R., Maddala,A., Meade,C.D., Williams,L.D. *et al.* (2021) R2DT is a framework for predicting and visualising RNA secondary structure using templates. *Nat. Commun.*, **12**, 3494.
14. Consortium,UniProt (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
15. Ochoa,D., Hercules,A., Carmona,M., Suveges,D., Gonzalez-Uriarte,A., Malangone,C., Miranda,A., Fumis,L., Carvalho-Silva,D., Spitzer,M. *et al.* (2021) Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res.*, **49**, D1302–D1310.
16. Laskowski,R.A., Stephenson,J.D., Sillitoe,I., Orengo,C.A. and Thornton,J.M. (2020) VarSite: disease variants and protein structure. *Protein Sci. Publ. Protein Soc.*, **29**, 111–119.
17. consortium,PDBe-KB (2020) PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res.*, **48**, D344–D353.
18. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
19. Iudin,A., Korir,P.K., Salavert-Torres,J., Kleywegt,G.J. and Patwardhan,A. (2016) EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods*, **13**, 387–388.
20. Cezard,T., Cunningham,F., Hunt,S.E., Koylass,B., Kumar,N., Saunders,G., Shen,A., Silva,A.F., Tsukanov,K., Venkataraman,S.

et al. (2022) The European Variation Archive: a FAIR resource of genomic variation for all species. *Nucleic Acids Res.*, **50**, D1216–D1220.
21. Bairoch,A. (2018) The Cellosaurus, a Cell-Line Knowledge Resource. *J. Biomol. Tech. JBT*, **29**, 25–38.
22. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
23. Robinson,J., Guethlein,L.A., Maccari,G., Blokhuis,J., Bimber,B.N., de Groot,N.G., Sanderson,N.D., Abi-Rached,L., Walter,L., Bontrop,R.E. *et al.* (2018) Nomenclature for the KIR of non-human species. *Immunogenetics*, **70**, 571–583.
24. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A. *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
25. Lawson,C.L., Patwardhan,A., Baker,M.L., Hryc,C., Garcia,E.S., Hudson,B.P., Lagerstedt,I., Ludtke,S.J., Pintilie,G., Sala,R. *et al.* (2016) EMDataBank unified data resource for 3DEM. *Nucleic Acids Res.*, **44**, D396–D403.
26. Schoch,C.L., Ciufo,S., Domrachev,M., Hotton,C.L., Kannan,S., Khovanskaya,R., Leipe,D., Mcveigh,R., O'Neill,K., Robbertse,B. *et al.* (2020) NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database J. Biol. Databases Curation*, **2020**, baaa062.
27. Liu,C., Shi,W., Becker,S.T., Schatz,D.G., Liu,B. and Yang,Y. (2021) Structural basis of mismatch recognition by a SARS-CoV-2 proofreading enzyme. *Science*, **373**, 1142–1146.
28. Spratt,A.N., Kannan,S.R., Woods,L.T., Weisman,G.A., Quinn,T.P., Lorson,C.L., Sönnerborg,A., Byrareddy,S.N. and Singh,K. (2021) Evolution, correlation, structural impact and dynamics of emerging SARS-CoV-2 variants. *Comput. Struct. Biotechnol. J.*, **19**, 3799–3809.
29. Alsulami,A.F., Thomas,S.E., Jamasb,A.R., Beaudoin,C.A., Moghul,I., Bannerman,B., Copoiu,L., Vedithi,S.C., Torres,P. and Blundell,T.L. (2021) SARS-CoV-2 3D database: understanding the coronavirus proteome and evaluating possible drug targets. *Brief. Bioinform.*, **22**, 769–780.
30. Banerjee,S., Seal,S., Dey,R., Mondal,K.K. and Bhattacharjee,P. (2021) Mutational spectra of SARS-CoV-2 orf1ab polyprotein and signature mutations in the United States of America. *J. Med. Virol.*, **93**, 1428–1435.
31. Yashvardhini,N., Jha,D.K. and Bhattacharya,S. (2021) Identification and characterization of mutations in the SARS-CoV-2 RNA-dependent RNA polymerase as a promising antiviral therapeutic target. *Arch. Microbiol.*, **203**, 5463–5473.